UNIVERSITI TUNKU ABDUL RAHMAN

Department of Mathematics and Actuarial Science

CONTENTS

6 Binary Variables and Logistic Regression 2

6.1 Probability distributions 2
6.2 Generalized linear models 4
6.3 Models for Binary Data 6
6.4 General logistic regression model . 24
6.5 Goodness of fit statistics 30
6.6 Residuals 35
6.7 Odds ratios and prevalence ratios . 38

6 Binary Variables and Logistic Regression

6.1 Probability distributions

We consider generalized linear models in which the outcome variables are measured on a binary scale. For example, the responses may be alive or dead, or present or absent. **Success** and **failure** are used as generic terms of the two categories.

First, we define the binary random variable

$$Z = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases}$$

with probabilities

$$P(Z = 1) = \pi$$

and

$$P(Z = 0) = 1 - \pi,$$

which is the Bernoulli distribution, Bernoulli(π).

If there are n such random variables $Z_1, \ldots, Z_n$ which are independent with $P(Z_j = 1) = \pi_j$, then their joint probability is

$$\prod_{j=1}^{n} \pi_j^{z_j}(1-\pi_j)^{1-z_j} = \exp\left[\sum_{j=1}^{n} z_j \ln\frac{\pi_j}{1-\pi_j} + \sum_{j=1}^{n}\ln(1-\pi_j)\right]$$

which is a member of the exponential family.

Next, for the case where the $\pi_j's$ are all equal, we can define

$$Y = \sum_{j=1}^{n} Z_j$$

so that Y is the number of successes in n "trials." The random variable Y has the Binomial distribution, $B(n, \pi)$:

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, y = 0, 1, \ldots, n.$$

Finally, we consider the general case of N are independent random variables $Y_1, Y_2, \ldots, Y_N$ corresponding to the numbers of successes in N different subgroups or strata (Table 6.1).

Table 6.1 Frequencies for N Binomial distributions

	Subgroups			
	1	2	...	N
Successes	Y_1	Y_2	...	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Total	n_1	n_2	...	n_N

If $Y_i \sim B(n_i, \pi_i)$, the log-likelihood function is

$$l(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^{N}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\ln(1-\pi_i) + \ln\binom{n}{y_i}\right]$$

where $\boldsymbol{\beta} = [\pi_1, \ldots, \pi_N]^T$ and $\mathbf{y} = [y_1, \ldots, y_N]^T$.

6.2 Generalized linear models

We want to describe the proportion of successes, $P_i = \frac{Y_i}{n_i}$, in each subgroup in terms of factor levels and other explanatory variables which characterize the subgroup. As $E(Y_i) = n_i\pi_i$ and so $E(P_i) = \pi_i$ we model the probabilities π_i as $g(\pi_i) = x_i^T\boldsymbol{\beta}$ where x_i is a vector of explanatory variables (dummy variables for factor levels and

measured values for covariates), $\boldsymbol{\beta}$ is a vector of parameters and g is a link function.

The simplest case is the linear model

$$\pi = x_i^T \boldsymbol{\beta}$$

This is used in some practical applications, but it has the disadvantage that although π is a probability, the fitted values $x_i^T \boldsymbol{\beta}$ may be less than zero or greater than one.

To ensure that π is restricted to the interval [0,1] it is often modeled using a cumulative probability distribution

$$\pi = \int_{-\infty}^{t} f(s)ds$$

where $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$. The probability density function $f(s)$ is called the **tolerance distribution**.

6.3 Models for Binary Data

The aim is to describe the probability of "success", π, as a function of x; for example, $g(\pi) = \beta_1 + \beta_2 x$.

1. If the tolerance distribution $f(s)$ is the Uniform distribution on the interval $[c_1, c_2]$,

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1}, & c_1 \leq s \leq c_2 \\ 0, & \text{otherwise} \end{cases}$$

then π is cumulative

$$\pi = \int_{c_1}^{x} f(s)ds = \frac{x - c_1}{c_2 - c_1}, \text{for } c_1 \leq x \leq c_2.$$

This equation has the form $\pi = \beta_1 + \beta_2 x$, where $\beta_1 = \frac{-c_1}{c_2 - c_1}$ and $\beta_2 = \frac{1}{c_2 - c_1}$.

This linear model is equivalent to using the identity function as the link function g and imposing conditions on x, β_1 and β_2 corresponding to $c_1 \leq x \leq c_2$. In practice,this model is not widely used.

2. **Probit Model**

If the tolerance distribution $f(s)$ is the standard normal distribution,

$$f(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2}, -\infty < s < \infty$$

$$\pi = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left[\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Thus

$$\Phi^{-1}(\pi) = \frac{x-\mu}{\sigma} = \beta_1 + \beta_2 x$$

where $\beta_1 = -\frac{\mu}{\sigma}$ and $\beta_2 = \frac{1}{\sigma}$.

In this case, the link function is Φ^{-1}.

Probit models are used in several areas of biological and social sciences in which there are natural interpretations of the model,

3. **Logistic or Logit Model**

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2},$$

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}.$$

This gives the link function

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$$

- $\ln\left(\frac{\pi}{1-\pi}\right)$ is called the logit function
- $\frac{\pi}{1-\pi}$ is the odds

The logistic model is widely used for Binomial data and is implemented in many statistical programs.

4. **Extreme value distribution**

$$f(s) = \beta_2 \exp[(\beta_1 + \beta_2 s) - \exp(\beta_1 + \beta_2 s)]$$

$$\pi = 1 - \exp[-\exp(\beta_1 + \beta_2 x)]$$

$$g(\pi) = \ln[-\ln(1-\pi)] = \beta_1 + \beta_2 x$$

$\ln[-\ln(1-\pi)]$ is called the **complementary log-log function**.

Example 1.
Table below shows numbers of beetles dead after five hours of exposure to gaseous carbon disulphide at various concentrations.

Dose, x_i $\log_{10} CS_2 mgl^{-1}$	Number of beetles, n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

We begin by fitting the logistic model

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

So,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$$

and

$$\ln(1-\pi_i) = -\ln(1+\exp(\beta_1 + \beta_2 x_i)$$

and

$$l = \sum_{i=1}^{N}\left[y_i(\beta_1+\beta_2 x_i) - n_i \ln(1+\exp(\beta_1+\beta_2 x_i) + \binom{n_i}{y_i}\right]$$

$$U_1 = \frac{\partial l}{\partial \beta_1} = \sum\left\{y_i - n_i\left[\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right]\right\} = \sum(y_i - n_i\pi_i)$$

$$U_2 = \frac{\partial l}{\partial \beta_2} = \sum\left\{y_i x_i - n_i x_i\left[\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right]\right\}$$

$$\mathcal{J} = \begin{bmatrix} \sum n_i\pi_i(1-\pi_i) & \sum n_i x_i \pi_i(1-\pi_i) \\ \sum n_i x_i \pi_i(1-\pi_i) & \sum n_i x_i^2 \pi_i(1-\pi_i) \end{bmatrix}$$

Maximum likelihood estimates are obtained by solving the iterative equation

$$\mathcal{J}^{(m-1)}\mathbf{b}^m = \mathcal{J}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

where the superscript (m) indicates the m^{th} approximation and $\mathbf{b}$ is the vector of estimates.

To obtained the MLE of $\boldsymbol{\beta}$, use R.

```
#To obtain the log-likelihood function
y=c(6,13,18,28,52,53,61,60)
n=c(59,60,62,56,63,59,62,60)
x=c(1.6907,1.7242,1.7552,1.7842,
1.8113,1.8369,1.8610,1.8839)
logit.Lik <- function(par, X, Y,n) {
  b0=par[1]
  b1=par[2]
  out=-sum(Y*(b0 + X*b1)-n*log(1+exp(b0 + X*b1)))
  return(out)
}
l = logit.Lik(c(0,0),x,y,n)
l
Output:
[1] 333.4038
```

The function above takes three arguments:

- **par** is a vector of parameter values (those of which likelihood has to be evaluated given the data);
- Y and X are two vectors of data, the dependent and the independent variable.


```
#Use the optim function to maximize
#the values of the parameters.
hand.logit =
optim(
par = c(0,0),# The starting values
fn = logit.Lik,# The function
X = x, # Vector of X
Y = y, # Vector of Y
n = n, #Vector of n
hessian=TRUE) #Ask to return the
               #"Hessian" matrix
beta = hand.logit$par
beta
Output:
[1] -60.71575  34.26947
```

The estimated value of the linear predictor is

$$\hat{\eta} = \hat{\beta}_1 + \hat{\beta}_2 x$$

The fitted values are

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)} = \frac{1}{1 + \exp[-(\hat{\beta}_1 + \hat{\beta}_2 x_i)]}$$

and

$$\hat{y}_i = n_i \hat{\pi}_i.$$

```
pih = exp(b1+b2*x)/(1+exp(b1+b2*x))
yh = n*pih
fitted values = cbind(pih,yh)
fittedvalues
output:
             pih        yh
[1,] 0.05861542  3.458310
[2,] 0.16405971  9.843583
[3,] 0.36216651 22.454324
[4,] 0.60535811 33.900054
[5,] 0.79519745 50.097439
[6,] 0.90324768 53.291613
[7,] 0.95520103 59.222464
[8,] 0.97905130 58.743078
```


For the final approximation, the estimated variance–covariance matrix for $\mathbf{b}$, $\mathcal{J}(\mathbf{b})^{-1}$ can be obtained from the inverse of the Hessian matrix.

The Hessian matrix is the matrix of second partial derivatives of the log-likelihood function with respect of the estimated parameters. Its inverse is the variance-covariance matrix of the ML estimates.

```
J = solve(hand.logit$hessian)
J
Output:
> J
          [,1]       [,2]
[1,]  26.83867 -15.081567
[2,] -15.08157   8.480252
```

$$\mathcal{J}(\mathbf{b})^{-1} = \begin{bmatrix} 26.840 & -15.082 \\ -15.0828.481 & \end{bmatrix}$$

The standard errors of the MLE are:

$S_{b_1} = \sqrt{26.84} = 5.18$ and $S_{b_2} = \sqrt{8.481} = 2.91$

```
SE = sqrt(diag(J))
SE
Output:
> SE
[1] 5.180605 2.912087
```

The deviance is

$$D = 2\sum\left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]$$

```
D = 2*sum(y*log(y/yh)+(n-y)&log((n-y-.5)/(n-yh-.5))
D
Output:
 11.5412
```

Under H_0, $D \sim \chi^2(6)$, since $D = 11.542 < \chi^2(6)_{.05} = qchisq(.95, 6) = 12.59$

We can also obtained the above estimation by using the glm function in R.


```
#1. data entry and manipulation
y=c(6,13,18,28,52,53,61,60)
n=c(59,60,62,56,63,59,62,60)
x=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839)
nmy=n-y
beetle.mat=cbind(y,nmy)

#2. logistic regression
glm1=glm(beetle.mat~x, family=binomial(link="logit"))
summary(glm1)

#3. To obtain pi hat i and y hat i
pih = fitted.values(glm1)
yh = n*pih
fv = cbind(pih,yh)
fv
Output:
Call:
glm(formula = beetle.mat ~ x, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
x             34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
```

```
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4

       pih        yh
1 0.05860103  3.457461
2 0.16402787  9.841672
3 0.36211901 22.451378
4 0.60531491 33.897635
5 0.79517177 50.095822
6 0.90323582 53.290913
7 0.95519611 59.222159
8 0.97904934 58.742961
```

Several alternative models can be fitted to the beetle mortality data. The results are shown in Table below. Among these models the extreme value model appears to fit the data best.

Comparison of observed numbers of killed with fitted values obtained form various dose-response models for the beetle mortality data.

Observed value of Y	Logistic model	Probit model	Extreme value model
6	3.46	3.36	5.59
13	9.84	10.72	11.2
18	22.45	23.48	20.9
28	33.90	33.82	30.3
52	50.10	49.62	47.7
53	53.29	53.32	54.1
61	59.22	59.66	61.1
60	58.74	59.23	59.9
D	11.23	10.12	3.45
b_1(s.e.)	–60.72(5.18)	–34.94(2.64)	–39.57(3.23)
b_2(s.e.)	34.27(2.91)	19.73(1.48)	22.04(1.79)

The correspond R are

```
#2. probit regression
glm2=glm(beetle.mat~x, family=binomial(link="probit"))
summary(glm2)
#To obtain pi hat i and y hat i
pih2 = fitted.values(glm2)
yh2 = n*pih2
fv2 = cbind(pih2,yh2)
```

```
fv2
Output:
Call:
glm(formula = beetle.mat ~ x, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
x             19.728      1.487   13.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.20  on 7  degrees of freedom
Residual deviance:  10.12  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4

        pih2       yh2
1 0.05691142  3.357774
2 0.17869349 10.721610
3 0.37874084 23.481932
4 0.60384830 33.815505
5 0.78754962 49.615626
6 0.90370973 53.318874
7 0.96233307 59.664650
8 0.98713279 59.227967
```



```
#3. Complementary log log
glm3=glm(beetle.mat~x, family=binomial(link="cloglog"))
summary(glm3)
#To obtain pi hat i and y hat i
pih3 = fitted.values(glm3)
yh3 = n*pih3
fv3 = cbind(pih3,yh3)
fv3

Output:
Call:
glm(formula = beetle.mat ~ x, family = binomial(link = "cloglog"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.80329  -0.55135   0.03089   0.38315   1.28883

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -39.572      3.240  -12.21   <2e-16 ***
x             22.041      1.799   12.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   3.4464  on 6  degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4

> #To obtain pi hat i and y hat i
> pih3 = fitted.values(glm3)
```

```
> yh3 = n*pih3
> fv3 = cbind(pih3,yh3)
> fv3
        pih3      yh3
1 0.09473644  5.58945
2 0.18801129 11.28068
3 0.33797124 20.95422
4 0.54231139 30.36944
5 0.75835580 47.77642
6 0.91767335 54.14273
7 0.98569859 61.11331
8 0.99912042 59.94723
```


Example 2.

You are given the following information for a fitted GLM:

Response variable: Occurance of accidents Response distribution: Binomial		
Parameter	df	β
Intercept	1	-2.417
Area	2	
Suburban	0	0
Urban	1	0.92
Rural	1	-1.2

(a) Determine the odds of an urban driver having an accident if the link selected is probit.

(b) Determine the odds of a rural driver having an accident if the link selected is complementary log-log.

(c) Determine the odds of a suburban driver having an accident if the link selected is logit.

6.4 General logistic regression model

The simple linear logistic model $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$ is a special case of the general logistic regression model

$$\text{logit}\pi = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x_i}^T\boldsymbol{\beta}$$

where $\mathbf{x}_i$ is a vector of continuous measurements corresponding to covariates and dummy variables corresponding to factor levels and $\boldsymbol{\beta}$ is the parameter vector.

Maximum likelihood estimates of the parameters $\boldsymbol{\beta}$, and consequently of the probabilities $\pi_i = g^{-1}(\mathbf{x_i}^T\boldsymbol{\beta})$, are obtained by maximizing the log-likelihood function

$$l(\boldsymbol{\pi}, \mathbf{y}) = \sum_{i=1}^{N}\left[y_i \ln \pi_i + (n_i = y_i)\ln(1-\pi_i) + \ln\binom{n}{y_i}\right]$$

The estimation process is essentially the same whether the data are grouped as frequencies for each covariate pattern (i.e., observations with the same values of all the explanatory variables) or each observation is coded 0 or 1 and

its covariate pattern is listed separately. If the data can be grouped, the response Y_i, the number of "successes" for covariate pattern i, may be modelled by the Binomial distribution. If each observation has a different covariate pattern, then $n_i = 1$ and the response Y_i is binary.

The deviance is

$$D = 2\sum\left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]$$

Goodness of fit can be assessed and hypotheses can be tested directly using the approximation

$$D \sim \chi^2(N - p)$$

where p is the number of parameters estimated and N the number of covariate patterns.

The estimation methods and sampling distributions used for inference depend on asymptotic results. For small studies or situations where there are few observations for each covariate pattern, the asymptotic results may be poor approximations.

Example 3. Embryogenic anthers
The data below are numbers y_{jk} of embryogenic anthers of the plant species Datura innoxia Mill. obtained when numbers n_{jk} of anthers were prepared under several different conditions. There is one qualitative factor with two levels, a treatment consisting of storage at 3°C for 48 hours or a control storage condition, and one continuous explanatory variable represented by three values of centrifuging force. We will compare the treatment and control effects on the proportions after adjustment (if necessary) for centrifuging force.

Embryogenic anther data

		Centrifuging force (g)		
Storage condition		40	150	350
Control	y_{1k}	55	52	57
	n_{1k}	102	99	108
Treatment	y_{2k}	55	50	50
	n_{2k}	76	81	90

We will compare three logistic models for π_{jk} the probability of the anthers being embryogenic, where $j = 1$ for the control group and $j = 2$ for the treatment group and $x_1 = \ln 140$, $x_2 = \ln 150$ and $x_3 = \ln 350$.

- Model 1: $logit\ \pi_{jk} = \alpha_j + \beta_j x_k$ (i.e., different intercepts and slopes);
- Model 2: $logit\ \pi_{jk} = \alpha_j + \beta x_k$ (i.e., different intercepts

but the same slope);

- Model 3: $logit\ \pi_{jk} = \alpha + \beta x_k$ (i.e., same intercept and slope).

These models were fitted by the method of maximum likelihood. The results are summarized below:

Maximum likelihood estimates and deviances for logistic models for the embryogenic anther data (standard errors of estimates in brackets).

Model 1	Model 2	Model 3
$a_1 = 0.234(0.628)$	$a_1 = 0.877(0.487)$	$a = 1.021(0.481)$
$a_2 - a_1 = 1.977(0.998)$	$a_2 - a_1 = 0.407(0.175)$	$b = -0.148(0.096)$
$b_1 = -0.023(0.127)$	$b = -0.155(0.097)$	
$b_2 - b_1 = -0.319(0.199)$		
$D_1 = 0.028$	$D_2 = 2.619$	$D_3 = 8.092$

To test the null hypothesis that the slope is the same for the treatment and control groups, we use $D_2 - D_1 = 2.591$. From the tables for the $\chi^2(1)$ distribution, the significance level is between 0.1 and 0.2, and so we could conclude that the data provide little evidence against the null hypothesis of equal slopes. On the other hand, the power of this test is very low and the estimates for Model 1 suggest that although the slope for the control group may be zero, the slope for the treatment group is negative. Comparison of the deviances from Models 2 and 3 gives a test for equality of the control and treatment effects after a common adjustment for centrifuging force: $D_3 - D_2 = 5.473$ which indicates that the storage effects are different. Obviously, Model 1 fits the data very well but this is hardly surprising since four parameters have been used to describe six data points—such "over-fitting" is not recommended!

R codes:

```
y = c(55,52,57,55,50,50)
n = c(102,99,108,76,81,90)
Storage = as.factor(c(1,1,1,2,2,2))
Force = log(c(40, 150,350,40, 150,350))
model1=glm(cbind(y,n-y)~Storage*Force,family=binomial(link="logit"))
summary(model1)
model2=glm(cbind(y,n-y)~Storage+Force,family=binomial(link="logit"))
summary(model2)
model3=glm(cbind(y,n-y)~Force,family=binomial(link="logit"))
summary(model3)

Output:
glm(formula = cbind(y, n - y) ~ Storage * Force, family = binomial(link

Deviance Residuals:
       1         2         3         4         5         6
 0.03611  -0.09370   0.05466   0.04305  -0.09855   0.05560

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.23389    0.62839   0.372   0.7097
Storage2        1.97711    0.99802   1.981   0.0476 *
Force          -0.02274    0.12685  -0.179   0.8577
Storage2:Force -0.31862    0.19888  -1.602   0.1091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

glm(formula = cbind(y, n - y) ~ Storage + Force, family = binomial(link

Deviance Residuals:
       1         2         3         4         5         6
```

```
-0.74964  -0.00509   0.72746   0.99006  -0.13512  -0.72744

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.87673    0.48701   1.800   0.0718 .
Storage2     0.40684    0.17462   2.330   0.0198 *
Force       -0.15459    0.09702  -1.593   0.1111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)


glm(formula = cbind(y, n - y) ~ Force, family = binomial(link = "logit")

Deviance Residuals:
      1        2        3        4        5        6
-1.5947  -0.8896  -0.2283   1.9610   0.8700   0.3204

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0213     0.4813   2.122   0.0338 *
Force        -0.1478     0.0965  -1.532   0.1255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  8.0916  on 4  degrees of freedom
AIC: 41.66

Number of Fisher Scoring iterations: 3
```


6.5 Goodness of fit statistics

Instead of using maximum likelihood estimation, we could estimate the parameters by minimizing the weighted sum of squares

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}$$

since $E(Y_i) = n_i\pi_i$ and $Var(Y_i) = n_i\pi_i(1-\pi_i)$.

This is equivalent to minimizing the Pearson chi-squared statistic

$$\chi^2 = \sum \left(\frac{o-e}{e}\right)^2$$

where o represents the observed frequencies, e represents the expected frequencies and summation is over all $2N$ cells. The reason is that

$$\begin{aligned}
\chi^2 &= \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i\pi_i(1-\pi_i)} \\
&= \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2(1-\pi_i) + [-(y_i - n_i\pi_i)]^2\pi_i}{n_i\pi_i(1-\pi_i)} \\
&= \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i}(1 - \pi_i + \pi_i) \\
&= S_w
\end{aligned}$$

When χ^2 is evaluated at the estimated expected frequencies, the statistic is

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

which is asymptotically equivalent to the deviances

$$D = 2\sum\left[y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]$$

The proof of the relationship between χ^2 and D uses the Taylor series expansion of $\ln(s/t)$ about $s = t$, namely,

$$s\ln\frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \cdots$$

Thus,

$$\begin{aligned} D &= 2\sum_{i=1}^{N}\left\{(\pi_i - n_i\hat{\pi}_i) + \tfrac{1}{2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + [(n_i - y_i)\right. \\ &\qquad \left. -(n_i - n_i\hat{\pi}_i)] + \tfrac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\hat{\pi}_i)]^2}{n_i - n_i\hat{\pi}_i} + \cdots \right. \\ &= \sum_{i=1}^{N}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \\ &= \chi^2 \end{aligned}$$

The asymptotic distribution of D, under the hypothesis that the model is correct, is $D \sim \chi^2(N-p)$, therefore, approximately $\chi^2 \sim \chi^2(N-p)$. The choice between D and χ^2 depends on the adequacy of the approximation to the $\chi^2(N-p)$ distribution. There is some evidence to suggest that χ^2 is often better than D because D is unduly influenced by very small frequencies. Both the approximations are likely to be poor, if the expected frequencies are too small (e.g., less than 1).

Sometimes the log-likelihood function for the fitted model is compared with the log-likelihood function for a minimal model, in which the values π_i are all equal (in contrast to the saturated model which is used to define the deviance). Under the minimal model, $\tilde{\pi}_i = \frac{\sum y_i}{\sum n_i}$. Let $\hat{\pi}_i$ denote the estimated probability for Y_i under the model of interest (so the fitted value is $\hat{y}_i = n_i\hat{\pi}_i$). The statistic is defined by

$$\begin{aligned} \mathcal{C} &= 2[l(\tilde{\boldsymbol{\pi}};\mathbf{y}) - l(\hat{\boldsymbol{\pi}};\mathbf{y})] \\ &= 2\sum_{i=1}^{N}\left[y_i\ln\left(\tfrac{\hat{y}_i}{n_i\tilde{\pi}_i}\right) - (n_i - y_i)\ln\left(\tfrac{n_i - \hat{y}_i}{n_i - n_i\tilde{\pi}_i}\right)\right] \end{aligned}$$

From the previous results, the approximate sampling distribution for $\mathcal{C}$ is $\chi^2(p-1)$ if all the p parameters except the intercept term β_1 are zero. Otherwise $\mathcal{C}$ will have a non-central distribution. Thus $\mathcal{C}$ is a test statistic for the hypothesis that none of the explanatory variables is needed for a parsimonious model. $\mathcal{C}$ is sometimes called the likelihood ratio chi-squared statistic.

By analogy with R^2 for multiple linear regression, another statistic sometimes used is

$$pseudo\ R^2 = \frac{l(\tilde{\boldsymbol{\pi}};\mathbf{y}) - l(\hat{\boldsymbol{\pi}};\mathbf{y})}{l(\tilde{\boldsymbol{\pi}};\mathbf{y})}$$

which represents the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared with the minimal model. This statistic

is produced by some statistical programs as a measure of goodness of fit. As for R^2, the sampling distribution of *pseudo* R^2 is not readily determined (so $p-$values cannot be obtained), and it increases as more parameters are added to the model. Therefore, various modifications of *pseudo* R^2 are used to adjust for the number of parameters.

The **Akaike information criterion** AIC and the **Schwartz or Bayesian information criterion** BIC are other goodness of fit statistics based on the loglikelihood function with adjustment for the number of parameters estimated and for the amount of data. These statistics are usually defined as follows:

$$AIC = -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2p$$

$$BIC = -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) + p \ln n$$

where p is the number of parameters estimated. The statistical software R uses this definition of AIC, for example.

Note that the statistics (except for *pseudo* R^2) discussed in this section summarize how well a particular model fits the data. So a small value of the statistic and, hence, a large p-value, indicates that the model fits well. These statistics are not usually appropriate for testing hypotheses about the parameters of nested models, but they can be particu-

larly useful for comparing models that are not nested.

For the logistic model for the beetle mortality example, the log-likelihood for the model with no explanatory variable is $l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) = -155.2002$ while $l(\hat{\boldsymbol{\pi}}; \mathbf{y}) = -18.7151$, so the statistic

- $\mathcal{C} = l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) - l(\hat{\boldsymbol{\pi}}; \mathbf{y}) = 2[-18.7151 - (-155.2002)] = 272.9702$
  This indicates that the slope parameter β_1 is definitely needed.
- *pseudo* $R^2 = \frac{l(\tilde{\boldsymbol{\pi}};\mathbf{y}) - l(\hat{\boldsymbol{\pi}};\mathbf{y})}{l(\tilde{\boldsymbol{\pi}};\mathbf{y})} = \frac{-155.2002-(-18.7151)}{-155.2002} = 0.8794$
  This indicates reasonable fit.
- $AIC = -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2p$
  $AIC_{Null} = 2(-155.2002) - 2(1) = 312.4004$
  $AIC_{proposed} = -2(-18.7151) - 2(2) = 41.4302.$
  This indicates the proposed model fits better than the null model.

R code to obtain loglikelihood value:

```
y=c(6,13,18,28,52,53,61,60)
n=c(59,60,62,56,63,59,62,60)
x=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839)
nmy=n-y
beetle.mat=cbind(y,nmy)
```

```
glm0=glm(beetle.mat~1, family=binomial(link="logit"))
glm1=glm(beetle.mat~x, family=binomial(link="logit"))
l0 = logLik(glm0)
l1 = logLik(glm1)
l0
l1

Output:
> l0
'log Lik.' -155.2002 (df=1)
> l1
'log Lik.' -18.71513 (df=2)
```

6.6 Residuals

For logistic regression there are two main forms of residuals corresponding to the goodness of fit measures D and χ^2. If there are m different covariate patterns, then m residuals can be calculated. Let Y_k denote the number of successes, n_k the number of trials and $\hat{\pi}_k$ the estimated probability of success for the k^{th} covariate pattern.

The **Pearson**, or **chi-squared, residual** is

$$\chi_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}, k = 1, \ldots, m$$

and

$$\sum_{k=1}^{m} \chi_k^2 = \sum_{k=1}^{m} \frac{(y_k - n_k\hat{\pi}_k)^2}{n_k\hat{\pi}_k(1-\hat{\pi}_k)} = \chi^2$$

The **standardized Pearson residuals** are

$$r_{pk} = \frac{\chi_k}{\sqrt{1-h_k}}$$

where h_k is the leverage, which is obtained from the hat matrix.

Deviance residuals can be defined similarly,

$$d_k = sign(y_k - n_k\hat{\pi}_k)) \left\{ 2\left[y_k \ln\left(\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k) \ln\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k} \right) \right] \right\}^{\frac{1}{2}}$$

and

$$\sum_{k=1}^{m} d_k^2 = D$$

where the term $sign(y_k - n_k\hat{\pi}_k))$ ensures that d_k has the same sign as χ_k.

The standardized deviance residuals are defined by

$$r_{Dk} = \frac{d_k}{\sqrt{1-h_k}}$$

Pearson and deviance residuals can be used for checking the adequacy of a model. For example, they should be plotted against each continuous explanatory variable in the model to check if the assumption of linearity is appropriate and against other possible explanatory variables not included in the model. They should be plotted in the order of the measurements, if applicable, to check for serial correlation. Normal probability plots can also be used because the standardized residuals should have, approximately, the standard Normal distribution $N(0, 1)$, provided the numbers of observations for each covariate pattern are not too small.

If the data are binary, or if n_k is small for most covariate patterns, then there are few distinct values of the residuals and the plots may be relatively uninformative. In this case, it may be necessary to rely on the aggregated goodness of fit statistics χ^2 and D and other diagnostics.

6.7 Odds ratios and prevalence ratios

Consider the case where the linear predictor has only a single regressor, so that the fitted value of the linear predictor at a particular value of x, say x_i , is

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \beta_1 x_i$$

The fitted value at $x_i + 1$ is

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \beta_1(x_i + 1)$$

and the difference in the two predicted values is

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1$$

Now $\hat{\eta}(x_i)$ is just the log-odds when the regressor variable is equal to x_i, and $\hat{\eta}(x_i + 1)$ is just the log-odds when the regressor is equal to $x_i + 1$. Therefore, the difference in the two fitted values is

$$\begin{aligned}\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) &= \ln(odds_{x_1+1}) - \ln(odds_{x_i}) \\ &= \ln\left(\frac{odds_{x_i+1}}{odds_{x_i}}\right) \\ &= \hat{\beta}_1\end{aligned}$$

If we take antilogs, we obtain the odds ratio

$$OR = \frac{Odds_{x_i+1}}{Odds_{x_i}} = e^{\hat{\beta}_1}$$

The odds ratio can be interpreted as the estimated increase in the probability of success associated with a one-unit change in the value of the predictor variable. In general, the estimated increase in the odds ratio associated with a change of d units in the predictor variable is $e^{d\hat{\beta}_1}$.

The interpretation of the regression coefficients in the multiple logistic regression model is similar to that for the case where the linear predictor contains only one regressor. That is, the quantity $e^{\hat{\beta}_j}$ is the odds ratio for regressor x_j, assuming that all other predictor variables are constant.

Example 4.
The data below concerning the proportion of coal miners who exhibit symptoms of severe pneumoconiosis and the number of years of exposure. The response variable of interest is the proportion of miners who have severe symptoms. A reasonable probability model for the number of severe cases is the binomial, so we will fit a logistic regression model to the data.

Years, x	5.8	15	21.5	27.5	33.5	39.5	46	51.5
Number of Miners, n	98	54	43	48	51	38	28	11
Number of severe cases, y	0	1	3	8	9	8	10	5

The R output is given below.

```
glm(formula = cases ~ x, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.79648    0.56859  -8.436  < 2e-16 ***
x            0.09346    0.01543   6.059 1.37e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 56.9028  on 7  degrees of freedom
Residual deviance:  6.0508  on 6  degrees of freedom
AIC: 32.877
```

(a) Fit a logistic regression model to the data. Use simple linear regression model as the structure for the linear predictor.

(b) Does the model deviance indicare that the the logistic regression from part (a) is adequate.

(c) Find the estimated probability of number of severe cases for a miner with 10 years of exposure and a sample size of 85.

(d) Interpret the slope β_2.

(e) Find the diviance residual for the third observation, d_3.